DOCUMENT RESUME

ED 340 745 TM 017 824

AUTHOR Wheeler, Patricia
TITLE The Relationship between Modified Angoff Knowledge Estimation Judgments and Item Difficulty Values for Seven NTE Specialty Area Tests.
PUB DATE Nov 91
NOTE 12p.; Paper presented at the Annual Meeting of the California Educational Research Association (70th, San Diego, CA, November 14, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Art Education; Biology; *Difficulty Level; Elementary Secondary Education; English; *Estimation (Mathematics); *Evaluators; Higher Education; Interrater Reliability; Language Tests; *Licensing Examinations (Professions); Mathematics Tests; Physical Education; Science Tests; Scoring; Social Studies; Spanish; *Teacher Certification; *Test Items
IDENTIFIERS Angoff Methods; California; *NTE Specialty Area Tests; Standard Setting; Subject Content Knowledge; Subject Specialists

ABSTRACT

The appropriateness of the Angoff method (W. H. Angoff, 1971) for setting standards on tests was studied. Evaluators (judges) from California school districts and teacher training institutions reviewed 15 NTE (National Teacher Examinations) Program Specialty Area Tests published by the Educational Testing Service for their appropriateness in teacher certification in California. In this paper, data are reported from more than 25 evaluators' reviews of NTE Specialty Area Tests in the following areas: (1) art education; (2) biology and general science; (3) English language and literature; (4) mathematics; (5) physical education; (6) social studies; and (7) Spanish. Each evaluator was assigned to review only one test in his or her subject matter specialty. Each multiple-choice test was reviewed by male and female representatives of two or more ethnic groups. Evaluators estimated the percentage of hypothetical candidates who would answer each item correctly. Results support the use of the modified Angoff standard-setting method and provide evidence that evaluators are not making arbitrary decisions in estimating item difficulty for barely acceptable candidates. To set appropriate standards, it is important to have qualified evaluators, good training, and adequate data and information for evaluators to use. Four data tables and a 13-item list of references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED340745

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PATRICIA WHEELER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE RELATIONSHIP BETWEEN MODIFIED ANGOFF KNOWLEDGE ESTIMATION JUDGMENTS AND ITEM DIFFICULTY VALUES FOR SEVEN NTE SPECIALTY AREA TESTS

Patricia Wheeler, Ph.D.
Senior Research Associate, CREATE, Western Michigan University and
Educational Research and Evaluation Consultant, Livermore, California

TM017824

BEST COPY AVAILABLE

The Relationship between Modified Angoff Knowledge Estimation Judgments and Item Difficulty Values for Seven NTE Specialty Area Tests

Patricia Wheeler, Ph. D.
Senior Research Associate, CREATE, Western Michigan University and
Educational Research and Evaluation Consultant, Livermore, California

A paper presented at the 70th annual meeting of the California Educational Research Association, San Diego, California, November 14, 1991.

Several methods have been used to set passing scores on examinations. No single method has been generally accepted as the best one due to the lack of research or to inconsistent findings when comparing the various methods for different purposes and with different types of examinee groups. However, the Angoff (1971) procedure (and variations of it) is the most widely used method for setting standards on teacher tests (Jaeger, 1990). This method relies on judges being able to estimate the difficulty level of each item for those examinees who would barely pass or who are marginal (Angoff, 1971; Livingston & Zieky, 1982).

Glass (1978) expressed concern about the arbitrariness of the Angoff and Nedelsky standard setting processes and whether judges can assess the difficulty of an item for minimally-acceptable candidates: "Can judges make such determinations consistently and reliably?" (p. 248). Mills and Melican (1987) were also concerned about judgmental approaches such as those applied in the Angoff method: "One problem associated with the use of item judgment methods has been the low to modest relationship of raters' perception of item difficulty to the actual item difficulty" (p. 1). This concern is addressed in the present study.

Several studies have examined the use of the Angoff and other methods from various perspectives. In response to Glass' article, Gross (1982) cited three medical education studies that reported positive relationships between p-values and item minimum pass indices derived using the Nedelsky method. Norcini et al. (1987) compared results of the Angoff method for three different groupings of judges--making judgments before, during and after a group meeting of the judges. Wheeler and Elias (1983) examined Angoff results derived from discussion group consensus versus a group working independently for the reading and mathematics sections of the California Basic Educational Skills Test (CBEST). DeMauro (1991) compared Angoff results for consistent and inconsistent judges for a modified version of an NTE Specialty Area Test. Cross et al. (1984) found positive correlations between knowledge estimation judgment scores and item difficulty levels (p-values) for two NTE Specialty Area Tests (Elementary Education and Mathematics). In their study, the Angoff, Ebel, and Nedelsky methods were examined to compare results when judges had only correct answers and when they had additional data on each item, including its p-value. Based on his review of studies on various standard setting methods, Jaeger (1988) found that the Angoff method tended to result in more reasonable standards for making certification decisions than other methods.

The purpose of this study is to further confirm the appropriateness of the Angoff method for setting standards on teacher tests. The focus of the present

TM017824

study is whether the judges' estimates of the percentage of minimally-acceptable examinees who would select the correct answer (knowledge estimates) are related to empirical item difficulty levels (p-values). Judges in this study reviewed 15 NTE Program Specialty Area Tests (published by Educational Testing Service, Princeton, New Jersey) for appropriateness for use in teacher credentialing in California. They provided data to the California Commission on Teacher Credentialing for use in setting passing scores on those tests that the Commission considered to be appropriate for use in the credentialing process (Wheeler et al., 1988).

Method

Tests. Although 15 NTE Specialty Area Tests were reviewed for the Commission, data are reported in this paper for those tests where more than 25 judges reviewed the test. These include the NTE Specialty Area Tests in the following areas:

- Art Education
- Biology and General Science
- English Language and Literature
- Mathematics
- Physical Education
- Social Studies
- Spanish

These tests are designed to measure knowledge and academic skills of potential teachers in the designated subject areas. All seven tests covered by this study are two-hour, multiple-choice tests of 120 to 160 items with five choices per item (except Spanish which has four choices per item).

Judges. Judges were selected by the Commission and were from school districts and teacher training institutions throughout the State of California. Tables 1 and 2 provide information on the demographic characteristics and professional background of the judges. Each test was reviewed by both male and female judges and by judges representing two or more ethnic groups. The judges were highly educated (most had graduate degrees) and very experienced in their fields (over 20 years, on the average, for each test).

The judges reviewed the tests during eight meetings in November and December of 1987 at sites throughout California. The meetings were led by members of the Commission staff following an outline developed jointly by the Commission and Educational Testing Service. One or more judges for each of the subject areas participated at each of the eight meeting. There were not separate meetings for each subject area. This approach ensured input, for each subject area, from a more diverse group of California educators in terms of regions of the State as well as number and types of work settings (e.g., public and private universities, high schools in small and large communities).

Procedures. Each judge was assigned to review only one test in his/her subject matter specialty. All judges for all subject areas were given the same instructions and used the same data collection form. Prior to reviewing test items, Commission staff discussed the purpose of the study and emphasized the need for

each judge, when making the knowledge estimation judgments, to focus on the minimally-acceptable candidate who should be considered for a credential, based on his/her performance on the NTE Specialty Area Test. They pointed out that this test is only one of several requirements that must be satisfied to be credentialed as a teacher in a subject area in California. They added that not all individuals taking these NTE Specialty Area Tests would be new teachers; some would be currently credentialed in another subject area and may have been teaching that other subject for several years. Judges were instructed to focus on the minimally-acceptable candidate teaching for the first time in the subject area of the test they were reviewing.

Judges were given a current form of the NTE Specialty Area Test. The correct answers were provided, but no other item information was given to them. Prior to doing the item review, judges carefully studied the test content outline to familiarize themselves with the knowledge areas and academic skills covered by the test. Judges spent about 45 minutes to one hour reviewing the outline and making comments. Next, judges were given some practice items and asked to judge what percent of a group of 100 minimally-acceptable candidates would mark the correct answer. Since the practice items were from several different subject areas, there were wide variations among the judges in different fields, but this process helped them focus on the minimally-acceptable group for their subject area.

Judges were provided with eight choices for estimating what percent of the hypothetical group would mark the answer correctly: 20, 30, 40, 50, 60, 70, 80, and 90+. Judges were asked to assume that all examinees attempted all questions. If such candidates were to guess randomly, the judges were instructed to mark 20 (since there were four or five possible answers per item, each item would be marked correctly by 20% of these examinees by chance). The estimation scale did not extend to 100 because it was assumed that, even on very easy items, some minimally-acceptable candidates would mark an incorrect answer due to misreading the item or misgridding the answer sheet. Although judges were provided with the correct answers for each question in the test book, they were instructed to read all distractors before making a choice because one or more of the distractors could be very attractive to the less knowledgeable candidates.

Results

In making a knowledge estimation (KE) judgment for each item, it was assumed that judges were considering primarily the difficulty of the item for minimally-acceptable candidates. Other factors that judges might have considered included the overall ability of such candidates, the importance of the knowledge area or academic skills measured by each item to the teacher training curriculum and to the beginning teacher's job in that subject area, and whether such candidates would have availed themselves of the opportunity to learn the knowledge area or acquire the academic skill measured by the item.

The mean knowledge estimation (KE) judgments for the California judges and the item difficulty levels (p-values), based on the examinees in the test analysis sample, are provided in Table 3. The test analysis sample is a national sample of examinees used by Educational Testing Service in reviewing test items and equating test forms. The Pearson product-moment correlations between

these values are shown in the right column; all are positive, at least .62, and significant at $p < .01$. These results indicate that the California judges' choices of knowledge estimation values on these seven NTE Specialty Area Tests are strongly related to the difficulty level of the items (as measured by the p-values for the test analysis sample). Even though these judges were highly educated and very experienced, they were able to judge how difficult these items would be for teacher credential candidates.

If we define the means of minimally-acceptable teacher candidates for the seven subject areas as the passing levels used in California, we find that these scores are higher than both the national mean scores and the test analysis sample mean scores for all seven tests (see Table 4). The passing levels correspond to "national-percent-below" figures ranging from 56 to 68, whereas the mean scores for the test analysis sample have comparable figures ranging from 41 to 52 percent below. Across all items, the mean KE judgments for minimally-acceptable teacher candidates are higher than the mean p-values of the test analysis sample for all but one test. However, we assumed that the test analysis sample would be comparable to a minimally-acceptable candidate group in California. Since both sets of scores have national percent below figures close to the middle of the score distribution, the relative magnitudes of the corresponding knowledge estimation judgments and p-values for each item would probably be similar. In the middle of the score distribution, test performance on a few items can have a larger effect on the obtained scaled scores and corresponding percent below figures than would be the case at either end of the distribution.

Conclusions and Recommendations

The results of this study support the use of the modified Angoff standard setting method and, in response to the concerns of Glass (1978) and of Mills and Melican (1987), evidence that judges are not making arbitrary decision when assessing item difficulty for minimally-acceptable candidates. Assuming that the rankings of p-values for the test analysis sample are similar to the rankings of p-values for a group of examinees scoring at or near the California passing levels, then judges are able to assess the difficulty level of each item for such examinees in making their knowledge estimation judgments for setting standards using the modified Angoff approach for these seven NTE Specialty Area Tests. The results are similar to those from other studies where item difficulty levels have been correlated with knowledge estimates collected from judges in a standards setting process.

Based on my experience conducting several standard setting studies using a modified Angoff approach and the results of this and other studies, several recommendations are provided below concerning the selection and training of judges, and the provision of data and information to judges. Judges selected to participate in standard setting studies should meet several criteria:

- they must be experienced (or knowledgeable) in the career area (or subject area) being assessed;
- they must have experience working as a mentor, supervisor, or trainer of the types of persons being assessed;
- they must be willing to take the task seriously and to use care in making judgments;

- they must be articulate so they can explain their judgments; and
- they must represent the key constituent groups to be assessed (e.g., gender, ethnic, work setting, linguistic).

Several key elements must be covered in the training of the judges. These include: the scope and purpose of the assessment; how assessment results will be used, both alone and in conjunction with other information about candidates; a description of the process of setting standards based on the collective judgments they provide; a discussion of the characteristics of a borderline or marginal candidate; and practice with several items during which judges share their judgments and are provided with p-values and other empirical data about the items from a defined group of examinees (e.g., the test analysis sample, all examinees tested last year in the state). Copies of relevant statutes and regulations, and information on other assessments and data that will be used in conjunction with this test should be available at the training session.

Providing the judges with some information, such as p-values for what is thought to be a borderline group of candidates, could have too much impact on them for their judgments to be considered independent, professional judgments. Judges should be provided with only the following test information in addition to the test items to be judged: a copy of the test content outline (or an abbreviated version of the test specifications), a copy of the answer document and of the oral and written directions provided to examinees. and a list of the correct answers. The judge is more apt to consider the distractors if the correct answer is given on a separate list rather than marked in the test book, and thus presumably is better able to judge how well borderline candidates will do on a given item. It should also be clear to judges how much time examinees have to complete the test and what instructions they are given about guessing. Some data should be available for the judges, but not distributed to them. These can include p-values, point biserials, and the percentage of examinees in each quartile or quintile selecting each answer option for each item. The rationales for correct answers on inferential or judgmental items should also be available to judges who have questions about specific items.

To set appropriate standards it is important to have qualified judges, good training, and adequate data and information for the judges to use. However, it is important not to provide information or data (such a p-values for what is thought to be a borderline group of examinees) that will lead the judges to make certain decisions for each item and for the test as a whole. Such a procedure might lead to more reliable judgments of passing scores, but the validity of such judgments may be questionable.

This study was done while the author was a Program Administrator at Educational Testing Service in Emeryville, California.

Table 1

Demographic Characteristics of the Judges

	Number of Judges	Gender		Ethnic Group					
		Female	Male	Asian	Black	Hispanic	Native Amer.	White	Other, No Resp.
Art Education	37	20	17	1	1	1	1	33	0
Biology and General Science	34	8	26	0	3	0	0	30	1
English Language and Literature	36	23	13	1	5	0	2	27	1
Mathematics	37	14	23	1	4	0	0	32	0
Physical Education	27	16	11	1	2	1	1	22	0
Social Studies	34	9	25	2	3	1	0	28	0
Spanish	31	19	12	0	0	9	1	17	4

Table 2

Professional Background of the Judges

	Number of Judges	Current Position			Education			Average Years of Professional Experience
		K-12 Tchr.	College Prof., Tchr. Trainer	Other	AB	MA	PhD	
Art Education	37	12	17	8	8	18	11	22.8
Biology and General Science	34	15	16	3	3	14	17	23.8
English Language and Literature	36	14	18	4	7	11	18	23.3
Mathematics	37	15	17	5	6	16	15	24.4
Physical Education	27	10	16	1	4	12	11	24.0
Social Studies	34	17	13	4	1	19	14	23.9
Spanish	31	16	15	0	8	8	15	21.1

Table 3

Knowledge Estimation and Item Difficulty Results by Test

	Number of Scored Items	Knowledge Estimation (KE)			Item Difficulty (PV)			Correlation *
		Number of Judges	Estimated Mean % Correct	Standard Deviation	Number of Examinees	Empirical Mean % Correct	Standard Deviation	$r_{ke, pv}$
Art Education	144	37	62	11	520	59	25	+.79
Biology and General Science	158	34	66	8	1,165	58	18	+.62
English Language and Literature	150	36	59	10	905	63	21	+.78
Mathematics	120	37	68	8	2,435	44	23	+.80
Physical Education	149	27	58	10	1,225	53	23	+.68
Social Studies	150	34	61	10	2,175	57	21	+.73
Spanish	160	31	64	9	280	63	18	+.71

* All correlations are significant at $p < .01$.

Table 4

Comparison of National Means, Test Analysis Sample, and California Passing Levels

Test	National Sample		Test Analysis Sample		Calif. Passing Level	
	Mean Score*	SD*	Mean Score**	Nat'l Percent Below*	Passing Score**	Corresponding Nat'l Percent Below*
Art Education	597	85	607	50	620	56
Biology and General Science	633	79	627	43	680	67
English Language and Literature	595	82	598	46	620	56
Mathematics	587	80	577	44	630	68
Physical Education	613	75	606	41	640	59
Social Studies	592	84	596	48	630	63
Spanish	597	97	611	52	650	65

* From Educational Testing Service, 1987.

** From Wheeler, et al., 1988.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.)* (pp. 508-600). Washington, DC: American Council on Education.

Cross, L. H., Impara, J. C., Frafy, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21*(2), 113-129.

DeMauro, G. (1991). *Setting standards on teacher assessments by maximizing the accuracy of Angoff judgments.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Educational Testing Service. (1987). *Interpreting scores on the NTE Specialty Area Tests, 1987-88.* Princeton, NJ: Author.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*(4) 237-261.

Gross, L. J. (1982). Standards and criteria: A response to Glass' criticism of the Nedelsky technique. *Journal of Educational Measurement, 19*(2), 159-162.

Jaeger, R. M. (1988). *Establishing standards on tests used for certification of education personnel: Validity issues.* Paper presented at the annual meeting of the American Psychological Association, Atlanta.

Jaeger, R. M. (1990). Setting standards on teacher certification tests. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary teachers* (pp. 295-321). Newbury Park, CA: SAGE Publications.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Mills, C. N., & Melican, G. J. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores* (Research Report RR-87-14). Princeton, NJ: Educational Testing Service.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24*(1), 56-64.

Wheeler, P., & Elias, P. (1983). *California Basic Educational Skills Test: Field test and validity study report.* Berkeley, CA: Educational Testing Service.

Wheeler, P., Hirabayashi, J. B., Martinson, J., & Watkins, R. W. (1988). *A study of the appropriateness of fifteen NTE Specialty Area Tests for use in credentialing in the state of California.* Emeryville, CA: Educational Testing Service. (A report prepared for the California Commission on Teacher Credentialing)